Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation

Year: 2023 | Citations: 484 | Authors: Lijun Yu, José Lezama, N. B. Gundavarapu, Luca Versari, Kihyuk Sohn

Abstract

While Large Language Models (LLMs) are the dominant models for generative tasks in language, they do not perform as well as diffusion models on image and video generation. To effectively use LLMs for visual generation, one crucial component is the visual tokenizer that maps pixel-space inputs to discrete tokens appropriate for LLM learning. In this paper, we introduce MAGVIT-v2, a video tokenizer designed to generate concise and expressive tokens for both videos and images using a common token vocabulary. Equipped with this new tokenizer, we show that LLMs outperform diffusion models on standard image and video generation benchmarks including ImageNet and Kinetics. In addition, we demonstrate that our tokenizer surpasses the previously top-performing video tokenizer on two more tasks: (1) video compression comparable to the next-generation video codec (VCC) according to human evaluations, and (2) learning effective representations for action recognition tasks.